Automatic Analysis of Multimodal Group Actions in Meetings

Iain McCowan, *Member, IEEE,* Daniel Gatica-Perez, *Member, IEEE,* Samy Bengio, *Member, IEEE,* Guillaume Lathoud, *Student Member, IEEE* Mark Barnard, *Student Member, IEEE,* Dong Zhang.
{mccowan,gatica,bengio,lathoud,barnard,zhang}@idiap.ch

***Abstract*— This paper investigates the recognition of group actions in meetings. A framework is employed in which group actions result from the interactions of the individual participants. The group actions are modelled using different HMM-based approaches, where the observations are provided by a set of audio-visual features monitoring the actions of individuals. Experiments demonstrate the importance of taking interactions into account in modelling the group actions. It is also shown that the visual modality contains useful information, even for predominantly audio-based events, motivating a multimodal approach to meeting analysis.**

Index Terms

Statistical models, Multimedia applications and numerical signal processing, Computer conferencing, Asynchronous interaction.

I. INTRODUCTION

Automatic analysis of meetings is an emerging domain for the research of a diverse range of speech, vision and multimodal technologies. Sample applications include structuring, browsing and querying of meeting databases, and facilitation of remote meetings.

Speech is the predominant modality for communication in meetings, and speech-based processing techniques, including speech recognition, speaker identification, topic detection, and dialogue modelling, are being actively researched in the meeting context [1], [2], [3], [4]. Visual processing, such as tracking people and their focus of attention, has also been examined in [5], [6]. Beyond this work, a place for analysis of text, gestures, and facial expressions, as well as many other audio, visual and multimodal processing tasks can be identified within the meeting scenario.

While important advances have been made, to date most approaches to automatic meeting analysis have been limited to the application of known technologies to extract information from individual participants (e.g. speech, gaze, identity, etc). Such a perspective overlooks the potential for defining new tasks based on the group nature of meetings. While producing accurate speech transcripts, identifying participants, and recognising visual gestures are all important tasks, one of the ultimate goals of automatic meeting analysis is the summarisation of the meeting into a series of high-level agenda items. Such a summarisation at the meeting level should reflect the action of the group as a whole, rather than simply actions of individual participants. Intuitively, the true information of meetings is created from interactions between participants: the whole is greater than the simple sum of the parts.

The automatic analysis of people interaction constitutes a rich research area. In domains other than meetings, there is growing interest in the automatic understanding of group behaviour, where the interactions are defined by individuals playing and exchanging both similar and complementary roles (e.g. a handshake, a dancing couple, or a children's game) [7], [8], [9], [10], [11]. Most of the previous work has relied on visual information and statistical models, and studied three specific scenarios: surveillance in outdoor scenes [10], [11], workplaces [8], [9], and indoor group entertainment [7]. In most cases, the interactions are composed of problem-dependent "primitive" tasks of various degrees of complexity performed by each individual, and selected from small sets of actions that are intuitively relevant. The main hypothesis in each of these cases is that the behaviour of people during an interaction is constrained by the behaviour of the others, so modelling such constraints amounts to modelling the interactions.

While little work has been done to date on automatic analysis of multimodal group interactions in meetings, group behaviour in meetings has been actively studied for over fifty years by social psychologists [12], [13], [14]. To develop technologies capable of analysing meetings automatically, much insight can be gained from familiarisation with this body of work. As a specific example, research has analysed the mechanisms and significance of turn-taking patterns in group discussions [15], [16], [17].

In this paper, we employ a statistical framework for automatic meeting analysis based on modelling interactions between participants (first presented in [18]). The actions of individual participants are first measured using a variety of audio-visual features. These multimodal feature sequences are then modelled in order to recognise actions belonging to the group as a whole (termed *meeting actions*). In particular, a set of meeting actions is defined based on turn-taking events. In experiments, we extract a range of audio-visual features from each participant (including speech activity, pitch, speaking rate, and head and hand blobs) and model the participant interactions using hidden Markov models (HMMs) [19]. The current experiments aim to investigate the multi-modal and group natures of the actions by using models that combine the streams of information (from audio, visual, or individuals) in different ways, including early integration HMMs, multi-

The authors are with the IDIAP Research Institute, Rue du Simplon 4, CP 592, CH-1920 Martigny, Switzerland.

stream HMMs [20], [21], coupled HMMs [22], and asynchronous HMMs [23].

As a background to the approach, Section II reviews related work from the field of social psychology. Section III then presents a computational framework for automatic meeting analysis based on the modelling of multimodal group actions. Experiments are presented in Section IV, and conclusions and future directions are given in Section V.

II. Meeting Analysis: A Social Psychology Perspective

While automatic meeting analysis is a recent research domain, a large body of literature on group interactions exists in the field of social psychology. This literature gives valuable insight into the nature and value of information present in meetings. In the following, we summarise aspects of the social psychology approach that are most relevant to the proposed computational perspective.

Social psychology concerns "the study of the manner in which the personality, attitudes, motivations, and behaviour of the individual influence and are influenced by social groups" [24]. Social psychology studies the above phenomena in a systematic manner and employs a variety of assessment methodologies, ranging from self-report measures and observational measures to physiological measures, among others [25]. Of these, we identify the *structured observational* approach (described below) as being of particular relevance to a computational framework. Further restricting our scope, we focus on studies of *small group discussions* [13], [17], as they relate well to the type of meetings we are currently investigating.

In *observational* approaches, group behaviour is measured by an observer/analyst. The analyst can observe either overtly or covertly, and may be external or internal to the group. Automatic analysis of meetings fits into this observational paradigm, where the machine functions as the observer/analyst.

More specifically, *structured* observational measures improve the objectivity of the analysis by defining a particular categorisation (the *coding system*) of group behaviour [25]. The categories in a given coding system can generally be considered as *mutually exclusive* (non-overlapping) and *exhaustive* (covering the entire meeting duration). In this way, the meeting can be annotated as a continuous sequence of these lexical labels. Structured approaches are commonly used when hypotheses about group behaviour can be probed by quantifying specific aspects of the group [25].

One distinction between different coding systems is that of *process* versus *task*. One process-based coding system is the Interaction Process Analysis (IPA) proposed by Bales [12], which is designed to measure how the group progresses through phases of communication, evaluation, control, decision, tension reduction and reintegration. The SYMLOG system (System of Multiple Level Observation of Groups) [26], is another process-based system based on attitudes of individuals within the group. The McGrath Task Circumplex [13] is an example of a task-based system. Its categories cover four broad task types - generate, choose, negotiate and execute - that translate into eight specific group tasks. An extension to the McGrath Task Circumplex was proposed in [27] to also include information sharing and gathering tasks. The lexica defined by the IPA and McGrath Task Circumplex coding systems are given in Table I.

System	Basis	Lexicon
IPA [12]	Process	shows solidarity shows tension release agrees gives suggestion gives opinion gives orientation asks for orientation asks for opinion asks for suggestion disagrees shows tension shows antagonism
McGrath [13]	Task	planning tasks creativity tasks intellective tasks decision-making tasks cognitive conflict tasks mixed-motive tasks contests/battles performances

TABLE I
ALTERNATIVE CODING SYSTEMS FOR GROUP DISCUSSIONS IN SOCIAL PSYCHOLOGY.

These coding systems are used to measure how individuals interact in a group, as well as how the group acts as a whole. Such group behaviours have direct relevance to potential applications, such as a meeting browser. To illustrate, Bales [12] gives a specific example of how the IPA categories could relate to potential meeting "agenda topics", and concludes that:

> "In brief, the functional problems of communication, evaluation, control, decision, tension reduction, and reintegration, have been separated out, enlarged into informal 'agenda topics' and made to form the skeleton of major events of the meeting." [12, p11].

Relating this to a computational framework, it is clear that automatic analysis of meetings can be considered a case of structured observational measurement. In this context, the meeting analysis task is defined as the recognition of a continuous, non-overlapping, sequence of lexical entries, analogous to the approach taken in speech or continuous gesture recognition [19], [28]. Each coding system provides an alternative lexicon of meeting events: the same meeting could be viewed from different perspectives by labelling according to a number of different coding systems in parallel.

One particular focus of group discussion research has been the 'morphology' of the group interaction, which investigates patterns of individuals' participation over time. Such analysis can give insight into issues such as interpersonal trust, cognitive load in interactions, and patterns of dominance and influence [14]. Recent work has shown that turn-taking patterns in meetings can be predicted [16] or simulated [15] using simple probabilistic models.

While it is evident that speaking turns are characterised predominantly by audio information, significant information is also present in non-verbal cues. Work has examined, for instance, how participants coordinate speaking turns using a variety of multimodal cues, such as gaze, speech back-channels, changes in posture, etc. [15], [16], [29]. Research has shown that in general, visual information can help disambiguate audio information [30], and that when the modalities are discrepant, participants appear to be more influenced by visual than by audio cues [14], [31].

Summarising the above discussion, the social psychological literature on group research provides valuable background information for automatic meeting analysis. In the current context, we have seen:

- that definition of a lexicon (coding system) of group events allows the interactions in meetings to be analysed in a systematic manner;
- that turn-taking behaviour provides a rich task for analysis; and
- that, while audio is the dominant modality in meetings, significant information is conveyed in the visual modality, motivating a multimodal approach.

III. Automatic Meeting Analysis: A Computational Framework

From the preceding discussion, we see that meetings can be analysed as a sequence of group actions that result from individuals interacting through a series of multimodal cues. Motivated by this view, this section describes a computational framework for automatic meeting analysis that involves three components: a set of multimodal group actions, a set of individual actions, and a model of the interactions.

A. Multimodal Group Actions

The first task in implementing such a framework, is to define a set of relevant group actions. As the actions belong to the meeting as a whole, rather than to any particular individual, we refer to them as *meeting actions*.

We model a meeting as a continuous sequence of exclusive events taken from the set of N meeting actions

$$V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}\,. \qquad (1)$$

We note that while the model of unambiguous, exclusive and exhaustive events provides a tractable computational framework, these assumptions do not always reflect reality. For instance, for events to be non-overlapping, it is implied that well-defined temporal boundaries exist. In reality, most events are characterised by soft (natural) transitions, and specifying their boundaries beyond a certain level of precision has little meaning. In addition, real events are not always perfectly unambiguous to observers (see e.g. [15], [27]). Nevertheless, such modelling inaccuracies are not necessarily limitations, depending on the particular application and assessment methodology.

While insight into the type of group actions present in meetings could be gained from the coding systems described in Table I, it is apparent that a computational framework requires a more constrained definition of meeting actions than that found in social psychology as recognition of the actions must be feasible given state-of-the-art technology.

As discussed in Section II, turn-taking provides a rich basis for analysing how people interact in group discussions. At its simplest level, segmenting a meeting into speaker turns is useful for structuring speech transcripts for browsing and retrieval. Analysis of speaker turns can also provide insight into the participants, such as their inherent latency in responding and degree of 'talkativeness', their role within a group, or their interest in particular topics [14], [15], [4].

Moving beyond simple speaker turns, turn-taking may be analysed at a higher-level by defining actions that may span several individual speaker turns, such as distinguishing between a series of monologues and a group discussion. Turns not based purely on speech, such as presentations, white-board usage or group note-taking, could also be defined if visual cues such as gaze and gestures were taken into account.

In this article, we propose an illustrative set of meeting actions based on high-level multimodal turns, including:

Monologue:
one participant speaks continuously without interruption,

Presentation:
one participant at front of room makes a presentation using the projector screen,

White-board:
one participant at front of room talks and makes notes on the white-board,

Discussion:
all participants engage in a discussion, and

(Group) Note-taking:
all participants write notes.

Specifically, in a meeting assumed to have four participants, we define a set of eight meeting actions to recognise as:

$$\begin{aligned} V = \{&\textit{`monologue1'}, \textit{`monologue2'}, \textit{`monologue3'}, \\ &\textit{`monologue4'}, \textit{`presentation'}, \textit{`white-board'}, \\ &\textit{`discussion'}, \textit{`note-taking'}\}. \end{aligned} \qquad (2)$$

These are all natural actions in which participants play and exchange similar, opposite, or complementary roles. For example, during a monologue, one person speaks to the group, while the other participants listen and direct their gaze towards the speaker or to their notes. During a discussion, multiple participants take relatively short turns at speaking, and more movement could be expected. In this set of actions, we define note-taking as a group event, in which the majority of participants take notes concurrently. Intuitively, it is expected that such an action would indicate periods where important information has been conveyed.

The value of segmenting a meeting according to this set of meeting actions is evident: it would, for example, facilitate browsing of a meeting archive by allowing the user to search for segments of most interest across the archive (such as presentations, or monologues by a particular person), and to quickly navigate between parts of the meeting for playback (see [32] for a simple demonstration of this for the corpus used in this paper). Experiments to recognise this set of meeting actions are presented in Section IV.

In a similar manner, other lexica of meeting actions could be defined to provide alternative views of a meeting. While actions should be non-overlapping within a given set of meeting actions, rich multi-layer views of meetings could be built by applying parallel sets of meeting actions to the same meeting. For example further lexica could be based on tasks (brainstorming, information sharing, decision making, etc), and the interest level of the group (high, neutral, low). Recent research in recognising emotion from speech [33], [34], recognising interest level from posture [35], recognising hot-spots (regions of high involvement or emphasis) in meetings [36], [37], [38], and detecting agreement and disagreement in meetings [39], suggests that the automatic recognition of such high-level concepts may become feasible.

B. Individual Actions

While many interesting and useful sets of meeting actions could be defined, whether or not a system can recognise them in practice depends on whether we can define and measure the constituent individual behaviour. For example, a presentation could intuitively be characterised by individual cues such as speech activity, location, and gaze. Similarly, brainstorming could involve short, approximately even-distributed speaker turns, individual note-taking, white-board use, and a characteristic set of speech keywords.

While the pertinence of these particular individual actions to the different meeting actions is somewhat speculative, it is clear from the above examples that many useful individual actions can be measured or recognised using state-of-the-art audio, visual and multimodal processing techniques.

These individual actions may be either fully recognised, or just measured. For example, individual actions including sitting, standing, raising hands, nodding and shaking heads, were recognised in [40]. While such recognised individual actions have value as annotations for browsing and indexing, direct measurements of the individual actions could be used as observable features when recognition of the group-level meeting actions is the goal. The experiments in this article investigate the latter approach. We denote an observation sequence $\mathbf{O}$ of T feature vectors as

$$\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_T), \tag{3}$$

where $\mathbf{o}_t$ is the vector of multimodal features at time t. Specifically, the experiments in this article investigate a set of audio-visual features, including: location-based speech activity; the pitch, energy and speaking rate of each participant; the location and orientation of each participant's head and hands; and the location of moving objects in the presentation and white-board regions. These features are described in detail in Section IV. We note that while the focus of the current article is to use these features directly to recognise group actions, we have also investigated recognition of individual actions based on this feature set in [41].

In general, such a set of features can be broken down into multiple feature streams, first according to participant i, and second according to modality m. We define the feature vector

$$o_t^{i,m} \in \mathbb{R}^{N_{i,m}}, \tag{4}$$

where $N_{i,m}$ is the number of features for individual i and modality m. We handle the case of participant-independent features (such as presentation area speech activity in this article), by replicating these for all values of i. To consider only features corresponding to a single individual, we define the notation

$$\mathbf{o}_t^{i,1:M} \triangleq \left(o_t^{i,1}, \dots, o_t^{i,M} \right), \tag{5}$$

where M is the number of modalities (here two, corresponding to audio and visual), and t the frame index. Similarly, to consider the feature vector for a single modality (across all individuals), we can define $\mathbf{o}_t^{1:I,m}$, where I is the number of participants, or to consider the set of all features $\mathbf{o}_t^{1:I,1:M}$.

Accordingly, we can define sequences of observations in the same way. For instance, $\mathbf{O}_l^{1:I,m}$, is the l^{th} sequence of observations represented by features of modality m, for all individuals.

C. Interaction Model

In order to model meeting actions, we propose to model the interactions between individuals. Considering these interactions as sequences of events, we can rely on the most successful approaches currently used to model temporal sequences of events, which are all based on a statistical framework. In this context, the general idea is to estimate, for each type of event $\mathbf{v}_j \in V$, the parameters θ_j of a distribution over corresponding sequences of observations $p(\mathbf{O}|\theta_j)$, where the sequence of observations $\mathbf{O}$ would correspond to the event $\mathbf{v}_j$. The most well-known solution to efficiently model such distributions is to use Hidden Markov Models (HMMs).

HMMs have been used with success for numerous sequence recognition tasks, including speech recognition [19], video segmentation [42], sports event recognition [43], and broadcast news segmentation [44]. HMMs introduce a state variable q_t and factor the joint distribution of a sequence of observations and the state using two simpler distributions, namely emission distributions $p(\mathbf{o}_t|q_t)$ and transition distributions $p(q_t|q_{t-1})$. Such factorisation yields efficient training algorithms such as the Expectation-Maximisation algorithm (EM) [45] which can be used to select the set of parameters θ_j^* of the model corresponding to event $\mathbf{v}_j$ to maximise the likelihood of L observation sequences as follows:

$$\theta_j^* = \arg\max_{\theta_j} \prod_{l=1}^{L} p(\mathbf{O}_l|\theta_j). \tag{6}$$

The success of HMMs applied to sequences of events is based on a careful design of sub-models (distributions) corresponding to lexical units (phonemes, words, letters, events). In the current framework, the lexical units are defined by the set of meeting actions $\mathbf{v}_j$, and a specific HMM will be created for each action $\mathbf{v}_j$. Given a training set of observation sequences representing meetings for which we know the corresponding labelling (but not necessarily the precise alignment), we create a new HMM for each sequence as the concatenation of sub-model HMMs corresponding to the sequence of meeting actions. This new HMM can then be trained using EM and will have the effect of adapting each sub-model HMM accordingly.

When a new sequence of observation features of a meeting becomes available, the objective is to obtain the optimal sequence of sub-model HMMs (representing meeting actions) that could have generated the given observation sequence. An approximation of this can be done efficiently using the well-known Viterbi algorithm [46].

While HMMs can be used to model various kinds of sequences of observations, several problems are in fact better described by multiple streams of observations, all corresponding to the same sequence of events [10], [20], [21], [22], [47]. This setup more closely corresponds to the case where each stream would represent the individual actions of a participant in a meeting, with the overall objective of analysing the interactions between individuals in terms of meeting actions.

Several solutions to the multiple stream setup have been proposed in the literature. The first and simplest one is to *merge* all observations related to all streams into one *large* stream (frame by frame), and to model it using a single HMM as explained above. This solution is often called *early integration*. Note that in some cases, when the streams represent information collected at different frame rates (such as audio and video streams for instance), up-sampling or down-sampling of the streams is first necessary in order to align the streams to a common frame rate.

Thus, using the notation introduced in Section III-B, the early integration solution is based on the creation of one model θ_j^* for each event $\mathbf{v}_j$ such that

$$\theta_j^* = \arg\max_{\theta_j} \prod_{l=1}^{L} p(\mathbf{O}_l^{1:I,1:M}|\theta_j). \tag{7}$$

A more complex option is the *multi-stream* approach [20]: in this case, each stream is modelled separately using its own HMM. For instance, if we consider the modalities as separate streams, we would create one model $\theta_{m,j}^*$ for each event $\mathbf{v}_j$ and modality m such that

$$\theta_{m,j}^* = \arg\max_{\theta_{m,j}} \prod_{l=1}^{L} p(\mathbf{O}_l^{1:I,m}|\theta_{m,j}). \tag{8}$$

Similarly, if we consider the individuals as separate streams, we would create one model $\theta_{i,j}^*$ for each event $\mathbf{v}_j$ and individual i such that

$$\theta_{i,j}^* = \arg\max_{\theta_{i,j}} \prod_{l=1}^{L} p(\mathbf{O}_l^{i,1:M}|\theta_{i,j}). \tag{9}$$

Then when a new meeting needs to be analysed, a special HMM is created, recombining all the single stream HMM likelihoods at various specific temporal points. Depending on these recombination points, various solutions appear. When the models are recombined after each state, the underlying system is equivalent to making the hypothesis that all streams are state-synchronous and independent of each other given the state. This solution can be implemented efficiently and has shown robustness to various stream-dependent noises. In the case of multiple modality streams, the emission probability of the combined observations of M streams in a given state of the model corresponding to event $\mathbf{v}_j$ at time t is estimated as:

$$p(\mathbf{o}_t^{1:I,1:M}|q_t) = \prod_{m=1}^{M} p(\mathbf{o}_t^{1:I,m}|q_t, \theta_{m,j}), \tag{10}$$

Similarly, in the case of multiple individual streams, the emission probability of the combined observations of I streams in a given state of the model corresponding to event $\mathbf{v}_j$ at time t is estimated as:

$$p(\mathbf{o}_t^{1:I,1:M}|q_t) = \prod_{i=1}^{I} p(\mathbf{o}_t^{i,1:M}|q_t, \theta_{i,j}). \tag{11}$$

One can see this solution as searching the best path into an HMM where each state i would be a combination of all states i of the single stream HMMs[1]. A more powerful recombination strategy enables some form of asynchrony between the states of each stream: one could consider an HMM in which states would include all possible combinations of the single stream HMM states. Unfortunately, the total number of states of this model would be exponential in the number of streams, hence quickly intractable. An intermediate solution, which we call *composite HMM*, considers all combinations of states in the same action only [48]. Hence, in this model, each action $\mathbf{v}_j$ HMM now contains all possible combinations of states of the corresponding action $\mathbf{v}_{m,j}$ of each stream HMM m. The total number of states remains exponential but is more tractable, when the number of states of each stream remains low (in our case around 3) as well as the number of streams (in our case, 2 or 4). The underlying hypothesis of this intermediate solution is that all streams are now action-synchronous instead of state-synchronous.

Multi-stream models are typically employed with separate streams for audio and visual features in multi-modal tasks [21], or for different frequency sub-bands in speech recognition [20]. In modelling group interactions however, the streams might instead represent the individual participants. This has the interesting advantage that the models could be trained for variable numbers of participants in meetings, and can even be used to decode meetings with a previously unseen number of participants. Moreover, the resulting decoding algorithm complexity is only linear in the number of participants.

Several other approaches to combine multiple streams of information have been proposed in the literature, but in general they suffer from an underlying training or decoding algorithm complexity which is exponential in the number of streams. For instance, *Coupled Hidden Markov Models* (CHMMs) [22], [49] can model two concurrent streams (such as one audio and one video stream) with two concurrent HMMs where the transition probability distribution of the state variable of each stream depends also on the value of the state variable of the other stream at the previous time step: more formally, let q and r be respectively the state variables of both streams, then CHMMs model transitions as follows: $p(q_t = i|q_{t-1} = j, r_{t-1} = k)$ and $p(r_t = i|r_{t-1} = j, q_{t-1} = k)$. Unfortunately,

[1] Note that this solution forces the topology of each single stream to be the same.

the exact training algorithm of such a model becomes quickly intractable when extended to more than 2 streams (which would be the case for meetings). An approximate algorithm which relaxes the requirement to visit every transition (termed the N-heads algorithm) was proposed in [49], and can be tractable for a small number of streams.

A more recent approach based on *Asynchronous Hidden Markov Models* (AHMMs) [23] models the joint probability of several streams by combining them in order to account for a possible asynchrony between them: it could be useful to temporarily stretch (or compress) a given stream with respect to the other ones. For instance, in a group action recognition task, an individual might start playing his/her role before the rest of the group. Being able to stretch the individual streams at specific points could yield performance improvement. While this approach has given promising results when there were only two streams, the currently proposed training algorithm quickly becomes intractable when extended to more than two streams. In the case of two modality streams (such as audio and video), an AHMM representing the event $\mathbf{v}_j$ models the joint distribution of the two streams by maximising the likelihood of L observation sequences as follows:

$$\theta_j^* = \arg\max_{\theta_j} \prod_{l=1}^{L} p(\mathbf{O}_l^{1:I,1}, \mathbf{O}_l^{1:I,2} | \theta_j). \qquad (12)$$

By introducing a state variable q_t (as for classical HMMs) and a synchronisation variable, τ_t, providing the alignment between the streams, one can factor the joint distribution into four simpler distributions, namely the transition distribution $p(q_t|q_{t-1})$, the joint emission distribution $p(\mathbf{o}_t^{1:I,1}, \mathbf{o}_t^{1:I,2}|q_t)$, the audio-only distribution $p(\mathbf{o}_t^{1:I,1}|q_t)$, and a distribution that models the fact that we should use the joint or the audio-only distribution at a given time $p(\text{emit}|q_t)$. Such factorisation yields efficient training and decoding algorithms when the number of streams is limited to two.

Apart from the models investigated in the current article, other models of interest include Layered HMMs and Dynamic Bayesian Networks (DBNs). Layered HMMs [47] are composed of layers, each of which takes its observation from the previous layer and generates the observation for the next layer. Experiments using Layered HMMs to recognise group actions from recognised individual actions (rather than directly from features, as in the current work) are presented in [41]. Dynamic Bayesian Networks (DBNs), a generalisation of HMMs, have also recently been applied with success to the same meeting recognition task described in this article, although only using the audio modality [50].

IV. Experiments

This section describes experiments to recognise multimodal meeting actions based on turn-taking events, as discussed in Section III-A. The following sub-sections describe the collection of a multi-modal database of these meeting actions, and then detail the experimental configuration and present results.

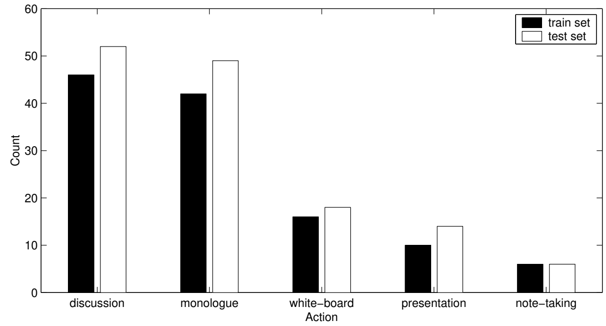


Fig. 1. Histogram showing occurrences of meeting actions in the train and test sets.

A. Data Collection

Data was collected in an instrumented meeting room which has dimensions 8.2m×3.6m×2.4m, and contains a 4.8m×1.2m meeting table. The room has been equipped with fully synchronised multi-channel audio and video recording facilities. For audio acquisition, 24 high quality miniature lapel microphones are simultaneously recorded at 48kHz with 24-bit resolution. The microphones are identical and are used both as close-talking lapel microphones attached to meeting participants, and in table-top microphone arrays. For video acquisition, three closed-circuit television cameras output PAL quality video signals, which are recorded onto separate MiniDV cassettes using three "video walkman" digital video tape recorders. Each camera is fitted with an adjustable wide-angle lens with a $38° - 80°$ field of view. Full details of the hardware setup are presented in [51].

A "scripted meeting" approach was taken to collect the required audio-visual data for the meeting action recognition experiments, to ensure adequate examples of all actions were included and also to facilitate annotation for training and testing.

An ergodic Markov model was used to generate meeting scripts. Each meeting action corresponded to a state in the Markov model with the self-loop transition probabilities governing the relative duration of each action. The transition probabilities were tuned by hand to ensure that the generated action sequences and durations were realistic. To illustrate this, the relative occurrences of different actions are shown in Figure 1 for the train and test sets (described below). On average, each meeting contained 5 actions. After generation of each meeting script, the action durations were normalised using a random time (in minutes) drawn from a $\mathcal{N}(5, 0.25)$ distribution, in order to constrain the total time to be approximately five minutes.

Two disjoint sets of eight meeting participants each were drawn from the local research staff population. For each set, thirty 4-person meeting scripts were generated as described above. The four participants for each meeting were chosen at random from the set of eight people. Every scripted meeting action in which a key role was played by a single participant (monologues, presentations, and white-boards) was then allocated at random to one of the four participants. Each meeting script was assigned a topic at random out of a small set of

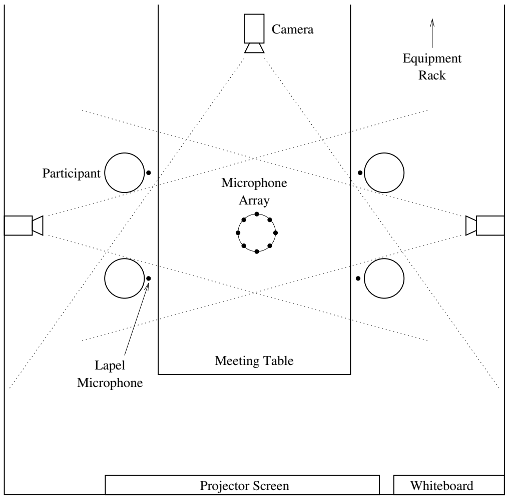


Fig. 2. Meeting recording configuration.

topics (e.g. my favourite movie). A dedicated timekeeper (off-camera) monitored the scripted action durations during meeting recording, and made silent gestures to prompt transitions between actions in the script. The behaviour of participants during actions was otherwise natural and unconstrained.

The meeting room configuration for the recordings is illustrated in Figure 2. Two cameras each acquired a front-on view of two participants including the table region used for note-taking. A third wide-view camera looked over the top of the participants towards the white-board and projector screen. The seating positions of participants were allocated randomly, with the constraint that participants who presented or used the white-board sat in one of the two seats closest to the front of the room (the latter was not exploited during analysis). All participants wore lapel microphones, and an eight-element circular equi-spaced microphone array of 20cm diameter was centrally located on the meeting table.

A total of 60 meeting recordings were collected (two participant sets, each having 30 meetings), resulting in approximately 5 hours of multi-channel, audio-visual meeting data. Each recording consists of three video channels, and twelve audio channels. The data is available for public distribution at [32].

B. Feature Extraction

Observation vectors are formed from a range of audio-visual features that measure the actions of individuals. These consist of :

Audio features :

Audio features were extracted from two different sources : the microphone array and the four lapels (one per participant).

From the microphone array signals, "speech activity" was estimated at 6 different locations : each of the four seats as well as the two locations corresponding to *'presentation'* and *'white-board'*. These locations were fixed 3-D vectors measured on-site, describing approximately where people would be standing or seated. "Speech activity" was computed as the Steered Response Power coming from each location using the SRP-PHAT measure [52], [53], which is a continuous, bounded value that indicates the activity of a particular location.

Using the streams of SRP-PHAT features, we were able to determine when each location was active. We thus obtained a speech/silence segmentation for each location, using a technique described in [54]. The segmentation was stored in order to compute the other features, but not present as a feature itself.

From each of the four lapel signals, we computed three additional acoustic features. The three acoustic features were energy, pitch and speaking rate, and were computed only on speech segments, setting a default value of zero on silence segments. Pitch was computed using the SIFT algorithm [55], speaking rate was obtained from a combination of estimators [56], and energy was calculated on each short-term (32 ms) Hamming-windowed segment. While these features were extracted from lapel signals in the current work, they could equally be extracted from the output of a microphone array beamformer for each participant (see [57], [58] for related research investigating developing beamforming and tracking algorithms for multiple people in a meeting room).

Finally, all 18 audio features were downsampled to match the 5 Hz rate chosen for video. Consecutive frames were merged, keeping the maximum value for each of the 6 SRP-PHAT features, and the median value for each of the 12 acoustic features.

Visual features :

Visual features were extracted using standard methods from image regions enclosing the seated participants (head and shoulders, the workspace at the table), and the white-board/presentation screen area.

For the cameras looking at people at the table, Gaussian Mixture Models (GMMs) of skin color in RGB space were used to extract head and hand/forearm blobs [59]. A 20-component GMM was estimated from the faces and arms of the people in the training set, which included caucasian, indian, and latin-american individuals. Skin pixels were then classified based on thresholding on the skin likelihood. A morphological postprocessing step was performed inside image regions enclosing typical head locations and the workspace to extract blobs.

For each person, the detected head blob was represented by the vertical position of its centroid (normalized by the average centroid computed over the meeting duration). Additionally, hand blobs were characterized by three features: the horizontal normalized centroid, the eccentricity, and the angle with respect to the horizontal [28]. Hand blob extraction and identification is especially difficult due to the free gesticulation patterns present in meetings. For instance, during a discussion the current speaker might introduce considerable self-occlusion while moving his hands (which might also occlude his face), while other participants might cross their arms or clasp their hands while listening. In this view, we opted to represent the hand blob information by using the described features for the right blob only (most participants in both training and test set are right-handed). Finally, a rough person motion feature was computed as the average of the individual motions of head and arms blobs, where motion

Fig. 3. Blob extraction in the multicamera meeting room. The top row of images shows a frame from each of the 3 cameras, and the bottom row shows the detected skin blobs (left and right) and moving blobs (centre).

	Modality		**Participants**	
Feature	**Audio** ($m = 1$)	**Visual** ($m = 2$)	**Individual** ($i = 1:4$)	**Other**
seat speech activity	✓		✓	
white-board speech activity	✓			✓
presentation speech activity	✓			✓
speech pitch	✓		✓	
speech energy	✓		✓	
speaking rate	✓		✓	
head blob vertical centroid		✓	✓	
hand blob horizontal centroid		✓	✓	
hand blob eccentricity		✓	✓	
hand blob angle		✓	✓	
combined motion		✓	✓	
white-board/presentation blob		✓		✓

TABLE II
BREAK-DOWN OF FEATURES ACCORDING TO STREAMS.

was computed as the centroid difference between consecutive frames. Note that while no tracking was performed at all, the tradeoff between the potential benefits for feature extraction, and the additional computational cost of a multi-part, multi-person tracker, remains to be seen.

For the wide-view camera, moving blobs were detected by background substraction and represented by their (quantised) horizontal position. A fixed background image was used, so errors in feature extraction due to sudden variations in the camera response occur, although not frequently. Adaptive background subtraction should improve robustness [60].

A typical result of blob extraction is shown in Figure 3 for the 3 different camera views. The final set of visual features consists of 21 features (5 for each seated participant, plus one from the whiteboard/screen camera).

This gives a total of 39 audio-visual features that were extracted at a frame-rate of 5 Hz.

C. Experimental Configuration

For the experiments, six different feature subsets were defined :

Audio-only:
all 18 audio features, trained according to Equation 8 with $m = 1$.

Visual-only:
all 21 visual features, trained according to Equation 8 with $m = 2$.

Individual participants (4):
12 (audio-visual) features. This consists of 9 person-specific features, plus the 3 other (participant-independent) features (replicated in each participant stream). Four separate streams trained according to Equation 9 with $i = 1 : 4$.

The specific features in these streams are summarised in Table II. We note that, the four streams for individual participants in fact correspond to the four different seating locations, and thus are independent of actual participant identities.

For the models, six HMM systems (mentioned in Section III-C) were used to combine these streams in different ways :

Early Integration:
single HMM trained on all 39 features, according to Equation 7.

Participant Multi-stream:
multi-stream HMM combining the 4 streams for

individual participants, with streams trained according to Equation 9. Two decoding schemes were investigated: state-level synchrony (Equation 11) and action-level synchrony (implemented using composite model within actions).

Participant Coupled:
coupled HMM combining the 4 streams for individual participants. The CHMM model was initialised using independently trained streams, and then retrained using an extension of the N-heads algorithm in [49] to an arbitrary number of streams. In decoding the action sequence, the streams were constrained by action-level synchrony.

Audio-Visual Multi-stream:
multi-stream HMM combining the audio-only and video-only streams, according to Equations 8 and 10. Two decoding schemes were investigated: state-level synchrony (Equation 10) and action-level synchrony (implemented using composite model within action models).

Audio-Visual Coupled:
coupled HMM combining audio-only and video-only streams, initialised and trained in a similar manner to the Participant CHMM above. In decoding the action sequence, the streams were constrained by action-level synchrony.

Audio-Visual Asychronous:
asychronous HMM combining the audio-only and video-only streams, according to Equation 12. To constrain complexity, the maximum allowed asynchrony between the streams was 2.2 seconds (compared to state duration of 0.2s and average action duration of 60s).

For all models, hyper-parameters (including number of emitting states per model (in range 1-3), number of GMM components per state (in range 1-10), and the insertion penalty for decoding) were selected using 5-fold cross-validation on the train set. For the AHMM, there were three distributions per state [23]: the audio distribution (GMM), the joint audio-visual distribution (GMM), and the visual emission probability distribution (binomial distribution). In this case, the audio stream was instead sampled at 10 Hz to better allow some form of asynchrony with the video stream.

All experiments were implemented using the Torch machine-learning library [61] (publicly available at [62]).

D. Results and Discussion

Results are presented in Table III in terms of the *action error rate* (AER) and the *frame error rate* (FER). The AER is equivalent to the word error rate used in automatic speech recognition (ASR). It is defined as the sum of insertion (extra actions recognised when no change occurred), deletion (actions omitted) and substitution (actions that occurred detected but labelled incorrectly) errors, divided by the total number of actions in the ground-truth, times one hundred. The use of the action error rate as a metric is appropriate when determining the correct sequence of events is more important than determining their precise temporal boundaries. This is the case here, due to the natural (ill-defined) transitions between the meeting actions [63]. The FER is the percentage of incorrectly labelled frames, and we include it here for two main reasons: it is necessary to verify that the temporal alignment of the recognised events is reasonable, and for reasons of statistical significance (see discussion of significance below). We note that the frame error rate enforces strict temporal boundaries, and is thus a harsh measure when such boundaries are inherently ill-defined, as is the present work.

Some results varied according to the random initialisation procedure in the EM-based training, which was exaggerated by the low number of training examples. Where this variation occurred, results presented are the mean and standard deviation over 10 runs.

As well as the results presented here, we note that the corpus can be browsed according to the resulting automatic transcriptions at [32].

1) Significance of Results: Due to the small number of actions present in the training and testing sets (around 140 in each), it is worth discussing the significance of these results. While standard deviations (where quoted) give an idea of how the various models are robust to initial conditions, statistical significance tests are often used to assess whether a model would be better than other ones on similar yet different test data. We have used a standard proportion test[2] [64], assuming a binomial distribution for the targets and using a normal approximation, which is often done in similar cases. In terms of action error rates, with 95% confidence, we cannot differentiate the 8 best models, namely audio-only, early integration, all audio-visual combinations, participant multi-stream with action-level synchrony, and participant coupled (note, these are also the 8 best in terms of FER). However, in terms of frame error rates, given the high number of test frames (more than 43000), all results are statistically significantly different from each other at a 95% level, hence for instance the best model (Audio-Visual Asynchronous) is statistically significantly better than the second best (Early Integration). While we consider the action error rate to be a more appropriate measure for these experiments, we therefore base the following discussion on the more reliable frame error rate results.

2) Single Streams: To help analyse these results, confusion matrices (from a randomly chosen single run) for the audio-only and visual-only streams are shown in Tables IV-V. It is clear that audio is the predominant modality for the set of meeting actions investigated here, being basically based on speaking turns, and this is reflected in the audio-only results. While less relevant information is present in the visual features, they are still able to give some discrimination between events. As would be expected, the visual features allow presentation and white-board to be recognised well. More interesting is the fact that they also give reasonable discrimination for discussion, which may be attributed to

[2]Note that action error rates are not really proportions/percentages since they can be greater than 100. Nevertheless, this test is often used to assess word error rates in ASR. On the other hand, this test is reasonable for frame error rates, which are indeed well defined proportions.

Model	Action Error Rate	Frame Error Rate
Audio-Only	15.8 (2.6)	11.2 (1.9)
Visual-Only	52.0 (2.8)	48.0 (2.7)
Individual Participants	39.6 (2.5)	32.2 (2.8)
Early Integration	8.9 (1.4)	10.0 (1.0)
Audio-Visual Multi-stream (state)	13.7	15.4
Audio-Visual Multi-stream (action)	13.0	16.3
Audio-Visual Coupled (action)	12.2	15.2
Audio-Visual Asynchronous	9.4 (0.3)	9.2 (0.1)
Participant Multi-stream (state)	19.1 (2.6)	18.4 (2.4)
Participant Multi-stream (action)	15.8 (1.4)	17.0 (1.1)
Participant Coupled (action)	13.6 (1.6)	16.9 (1.2)

TABLE III

ACTION ERROR AND FRAME ERROR RATES (IN PERCENT, LOWER IS BETTER) ON THE TEST SET WITH VARIOUS HMM ARCHITECTURES MODELLING MEETING ACTIONS. WHERE THE INITIALISATION PROCEDURE INTRODUCED VARIATION IN RESULTS, THE VALUES GIVEN ARE THE MEAN AND STANDARD DEVIATION (PARENTHESISED) OVER 10 RUNS. CONSTRAINTS ON SYNCHRONY (STATE-LEVEL OR ACTION-LEVEL) ARE INDICATED FOR APPROPRIATE MULTIPLE STREAM MODELS.

increased motion of participants. Here we see that neither modality in isolation is capable of distinguishing the note-taking periods, perhaps as it is jointly characterised by both audio silence and visual gestures.

Table VI shows that the single participant streams are able to give some discrimination between events, however as the actions essentially occur at the group level, the individual streams contain insufficient information to distinguish them reliably. In particular, the individual streams are not able to distinguish monologues well. This behaviour could be improved if accurate gaze features were used, as this should be a reliable indicator of silent participants' focus of attention (during others' monologues) [15].

3) Early Integration: Examining the different combination approaches, we note that early integration gives significantly better frame error rates than all approaches apart from the audio-visual AHMM. The improvement over the audio-only results comes mostly from the improved recognition of note-taking, as shown in the confusion matrix in Table VII. This result highlights the benefit of the multi-modal approach: while neither modality in isolation was able to reliably recognise note-taking, their combination achieves almost perfect results for this action. The other improvement we see over the audio-only results is a reduction in monologue and discussion insertion and deletion errors. The extra monologues in the audio-only results were mostly inserted in the middle of discussions, and so it is seen that the motion present in the video stream helps in discriminating discussion from monologues.

4) Audio-Visual Multi-stream, Coupled and AHMM: All models using separate audio and visual streams (multi-stream HMM, CHMM, AHMM) give good results in terms of the action error rate. However, we see from the frame error rate that only the AHMM system is significantly better than the audio-only stream in isolation. This demonstrates the importance of modelling the feature-level correlation between modalities, which is disregarded in the case of the multi-stream HMM and, to a lesser extent, the coupled HMM (which only models state-level correlation between streams). By comparing the systems with state-synchrony to those with action-synchrony, we see that there is no significant asynchrony between the audio and visual streams. This is also confirmed by the closeness of the results for the audio-visual AHMM and the early integration HMM.

5) Participant Multi-stream and Coupled: While the state-synchronous multi-stream combination of the four participant streams performs better than each stream in isolation, this is significantly lower than for the early integration approach. The action-synchronous multi-stream results demonstrate that a significant improvement can be achieved by allowing asynchrony between participants. While there is a small improvement using the coupled HMM over the multi-stream HMM, the performance is still lower than the early integration approach, highlighting the need to model feature-level correlation between participants.

E. Summary

Summarising the above discussion, we make a few observations based on these results:

1) There is benefit in a multi-modal approach to modelling group actions in meetings.
2) It is important to model the correlation between the behaviour of different participants.
3) There is no significant asynchrony between audio and visual modalities for these actions (at least within the resolution of the investigated frame rate).
4) There is evidence of asynchrony between participants acting within the group actions.

The above findings appeal to the intuition that individuals act in a group through both audio and visual cues which can have a causal effect on the behaviour of other group members. As a final remark, these results lead us to hypothesise that the AHMM with participant streams would provide a powerful model for group actions, highlighting the need to seek a tractable training algorithm for the case of multiple (> 2) streams, and more significant asynchrony ($>$ 2s).

	disc	mono1	mono2	mono3	mono4	note	pres	white	DEL
disc	44								7
mono1		10		1					1
mono2	1		10			1			
mono3				16					
mono4					10				1
note									5
pres							12		1
white							1	18	
INS		1	2		1				

TABLE IV

CONFUSION MATRIX OF RECOGNISED MEETING ACTIONS FOR AUDIO-ONLY, INCLUDING DISCUSSIONS (DISC), MONOLOGUES (MONO1-4), NOTE-TAKING (NOTE), PRESENTATIONS (PRES), AND WHITE-BOARDS (WHITE), AS WELL AS INSERTION ERRORS (INS) AND DELETION ERRORS (DEL). ZERO VALUES ARE REPRESENTED AS EMPTY CELLS. COLUMNS AND ROWS SHOW DESIRED AND OBTAINED LABELS, RESPECTIVELY.

	disc	mono1	mono2	mono3	mono4	note	pres	white	DEL
disc	30	3		3	1				12
mono1	6	1		2					5
mono2			2	1	1	1			8
mono3	1			2	1	1			8
mono4	2	2		1	3				5
note						1			3
pres							12		1
white							1	18	
INS	3								

TABLE V

CONFUSION MATRIX OF RECOGNISED MEETING ACTIONS FOR VIDEO-ONLY.

	disc	mono1	mono2	mono3	mono4	note	pres	white	DEL
disc	38	1		1					4
mono1	8		1		5	2			3
mono2	2	4	5	7					4
mono3				1					5
mono4		2		3					6
note		1				1			3
pres							12		1
white							1	18	
INS		1	1		2				

TABLE VI

CONFUSION MATRIX OF RECOGNISED MEETING ACTIONS FOR AN INDIVIDUAL PARTICIPANT.

	disc	mono1	mono2	mono3	mono4	note	pres	white	DEL
disc	49								3
mono1		11							
mono2			10						
mono3				15					2
mono4					7				4
note						5			1
pres							12		1
white							1	18	
INS		1							

TABLE VII

CONFUSION MATRIX OF RECOGNISED MEETING ACTIONS FOR THE EARLY INTEGRATION SYSTEM.

F. Application to Real Meeting Data

The meeting corpus for the above experiments was necessarily constrained to facilitate training and testing. To verify the robustness of the technique on natural data, a one-hour, four-participant real meeting was recorded for analysis. Features were extracted, and meeting actions were recognised using three of the best models for the differing numbers of streams, namely early integration, the state-synchronous multi-stream model for the audio-visual streams, and the coupled HMM for the 4 participant streams. The model parameters are the same ones used for the previous experiments, without any tuning.

To objectively assess the ability of the system to recognise the meeting actions, an effort was made to produce a ground-truth transcription of the meeting. In observing this data, however, it was apparent that in reality it is not obvious how to draw an absolute distinction between actions like monologues and discussions. We opted for the following approach for evaluation. Each sequence of recognised actions was verified by two independent observers not familiar with the system. The subjects played back the meeting recordings in real-time, and judged the correctness of each recognised action in the corresponding time interval, proposing a new action label if appropriate. Six subjects participated in the experiment. In a second step, a decision was taken by a third person (one of the authors) for those actions that were in disagreement among each pair of observers.

The classification results are shown in Table VIII. For all models, most of the difficulties, both for people and the automatic algorithms, arise from the ambiguity existing between actions originally defined as non-overlapping (e.g. between monologues and discussions, or due to the temporal co-occurrence of actions, like note-taking by one of the participants in the middle of a discussion).

While highlighting the difficulty and subjectivity of the task, this analysis also suggests that the system provides a segmentation that is reasonable to a human observer, and which thus has value for applications such as browsing and indexing. However, it is apparent that future research needs to address the ill-defined nature of some actions in real data.

V. Conclusions and Future Directions

In this paper we have presented an approach to automatic meeting analysis that considers a meeting as a sequence of group-level events, termed meeting actions. These meeting actions result from the interactions between individual participants, and are inherently multimodal in nature.

An illustrative set of meeting actions, based on high-level turn-taking behaviour, was defined. These actions were recognised in experiments using a range of audio-visual features extracted from each participant, and modelled using different HMM-based approaches. The best results were achieved by the audio-visual Asynchronous HMM system, which gave an action error rate of 8.9%, confirming the importance of modelling the interactions between individuals, as well as the advantage of a multimodal approach.

While the experiments in this article have shown the successful recognition of a set of turn-based meeting actions, there is much scope for future work to recognise other sets of high-level meeting actions, such as group level-of-interest. To achieve this goal, ongoing work is investigating richer feature sets (such as gaze, recognition of individual actions) and different means of modelling the multimodal interactions of participants. This will involve the collection of a larger, more natural, meeting corpus, as well as the development of more flexible assessment methodologies.

VI. Acknowledgements

The authors would like to acknowledge the invaluable advice of Jean Carletta (Human Communication Research Centre, Edinburgh University) regarding small group research in social psychology. We also acknowledge our colleagues at IDIAP for their assistance during the data collection, and the evaluation of the results in real meetings.

This work was supported by the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The work was also funded by the European projects "M4: MultiModal Meeting Manager" and "LAVA: Learning for Adaptable Visual Assistants", through the Swiss Federal Office for Education and Science (OFES).

References

[1] F. Kubala, "Rough'n'ready: a meeting recorder and browser," *ACM Computing Surveys*, no. 31, 1999.
[2] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc. of the Human Language Technology Conference*, (San Diego, CA), March 2001.
[3] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. IEEE ICASSP*, (Salt Lake City, UT), May 2001.
[4] S. Renals and D. Ellis, "Audio information access from meeting rooms," in *Proc. IEEE ICASSP 2003*, 2003.
[5] A. Waibel, T. Schultz, M. Bett, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang, "SMaRT:the Smart Meeting Room Task at ISL," in *Proc. IEEE ICASSP 2003*, 2003.
[6] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcasting system," in *Proc. ACM Multimedia Conference*, 2002.
[7] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schutte, and A. Wilson, "The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment," *PRESENCE: Teleoperators and Virtual Environments*, vol. 8, August 1999.
[8] N. Johnson, A. Galata, and D. Hogg, "The acquisition and use of interaction behaviour models," in *Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition*, June 1998.
[9] T. Jebara and A. Pentland, "Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour," in *Proc. International Conference on Vision Systems*, January 1999.
[10] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, August 2000.
[11] S. Hongeng and R. Nevatia, "Multi-agent event recognition," in *Proc. IEEE Int. Conference on Computer Vision*, (Vancouver), July 2001.
[12] R. F. Bales, *Interaction Process Analysis: A method for the study of small groups*. Addison-Wesley, 1951.
[13] J. E. McGrath, *Groups: Interaction and Performance*. Prentice-Hall, 1984.
[14] J. McGrath and D. Kravitz, "Group research," *Annual Review of Psychology*, vol. 33, pp. 195–230, 1982.
[15] E. Padilha and J. C. Carletta, "A simulation of small group discussion," in *EDILOG*, 2002.
[16] K. C. H. Parker, "Speaking turns in small group interaction: A context-sensitive event sequence model," *Journal of Personality and Social Psychology*, vol. 54, no. 6, pp. 965–971, 1988.
[17] N. Fay, S. Garrod, and J. Carletta, "Group discussion as interactive dialogue or serial monologue: The influence of group size," *Psychological Science*, vol. 11, no. 6, pp. 487–492, 2000.
[18] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interactions in meetings," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, April 2003.
[19] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.
[20] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, 2001.
[21] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, pp. 141–151, September 2000.
[22] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *Proceedings of IEEE CVPR97*, 1997.

Model	Number of recognised actions	Classification rate
Early Integration	36	88.8
Audio-Visual Multi-stream (state)	42	76.2
Participant Coupled (action)	46	84.8

TABLE VIII

ACTION CLASSIFICATION RATES (IN PERCENT, HIGHER IS BETTER) FOR THE THREE BEST HMM MODELS, ON A ONE-HOUR REAL MEETING. CONSTRAINTS ON SYNCHRONY (STATE-LEVEL OR ACTION-LEVEL) ARE INDICATED FOR APPROPRIATE MULTIPLE STREAM MODELS.

[23] S. Bengio, "An asynchronous hidden markov model for audio-visual speech recognition," in *Advances in Neural Information Processing Systems, NIPS 15* (S. Becker, S. Thrun, and K. Obermayer, eds.), MIT Press, 2003.

[24] "Merriam-webster online dictionary." http://www.m-w.com/.

[25] D. Forsyth, "Measurement in social psychological research," *http://www.people.vcu.edu/∼jforsyth/methods/measure.htm*, 2003.

[26] R. F. Bales and S. P. Cohen, *SYMLOG: A System for the Multiple Level Observation of Groups*. The Free Press, 1979.

[27] K. Ward, C. Marshall, and D. Novick, "Applying task classification to natural meetings," Tech. Rep. CS/E 95-011, Oregon Graduate Institute, 1995.

[28] T. Starner and A. Pentland, "Visual recognition of american sign language using HMMs," in *Proc. Int. Work. on Auto. Face and Gesture Recognition*, (Zurich), 1995.

[29] D. Novick, B. Hansen, and K. Ward, "Coordinating turn-taking with gaze," in *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP-96)*, 1996.

[30] R. Krauss, C. Garlock, P. Bricker, and L. McMahon, "The role of audible and visible back-channel responses in interpersonal communication," *Journal of Personality and Social Psychology*, vol. 35, no. 7, pp. 523–529, 1977.

[31] B. DePaulo, R. Rosenthal, R. Eisenstat, P. Rogers, and S. Finkelstein, "Decoding discrepant nonverbal cues," *Journal of Personality and Social Psychology*, vol. 36, no. 3, pp. 313–323, 1978.

[32] "IDIAP data distribution." http://mmm.idiap.ch/.

[33] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speech signals," in *Proc. Eurospeech*, (Geneva), Sept. 2003.

[34] V. Hozjan and Z. Kacic, "Improved emotion recognition with large set of statistical features," in *Proc. Eurospeech*, (Geneva), Sept. 2003.

[35] S. Mota and R. Picard, "Automated posture analysis for detecting learner's interest level," in *Proc. CVPR Workshop on Computer Vision and Pattern Recognition for Human Computer Interaction (CVPRHCI)*, (Madison Wisconsin,), June. 2003.

[36] B. Wrede and E. Shriberg, "Spotting hotspots in meetings: Human judgments and prosodic cues," in *Proc. Eurospeech*, (Geneva), Sept. 2003.

[37] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proc. ASRU*, (Virgin Islands), Dec. 2003.

[38] L. Kennedy and D. Ellis, "Pitch-based emphasis detection for characterization of meeting recordings," in *Proc. ASRU*, (Virgin Islands), Dec. 2003.

[39] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proc. HLT-NAACL Conference*, (Edmonton), May 2003.

[40] M. Zobl, F. Wallhoff, and G. Rigoll, "Action recognition in meeting scenarios using global motion features," in *Proceedings of the ICVS Workshop on Performance Evaluation of Tracking and Surveillance*, March 2003.

[41] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Modelling individual and group events in meetings: A two-layer hmm framework," IDIAP-RR 04-09, IDIAP, 2004.

[42] J. S. Boreczky and L. D. Wilcox, "A Hidden Markov Model framework for video segmentation using audio and image features," in *Proceedings of the ICASSP 98*, vol. 6, (Seattle), pp. 3741–3744, 1998.

[43] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with Hidden Markov Models," in *ICASSP*, 2002.

[44] S. Eickeler and S. Müller, "Content-based video indexing of TV broadcast news using Hidden Markov Models," in *Proceedings of the ICASSP 99*, (Phoenix), pp. 2997–3000, 1999.

[45] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[46] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, pp. 260–269, 1967.

[47] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for learning and inferring office activity from multiple sensory channels," in *Proceedings of the International Conference on Multimodal Interfaces (ICMI'02)*, October 2002.

[48] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing* (G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, eds.), MIT Press, 2004.

[49] M. Brand, "Coupled hidden markov models for modeling interacting processes," TR 405, MIT Media Lab Vision and Modeling, November 1996.

[50] A. Dielmann and S. Renals, "Dynamic bayesian networks for meeting structuring," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, (Montreal), May 2004.

[51] D. Moore, "The IDIAP smart meeting room," *IDIAP Communication 02-07*, 2002.

[52] J. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments*. PhD thesis, Brown University, Providence RI, USA, 2000.

[53] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays* (M. Brandstein and D. Ward, eds.), ch. 8, pp. 157–180, Springer, 2001.

[54] G. Lathoud, I. A. McCowan, and D. C. Moore, "Segmenting Multiple Concurrent Speakers Using Microphone Arrays," IDIAP-RR 21, IDIAP, Martigny, Switzerland, 2003. Submitted for publication.

[55] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, pp. 367–377, 1972.

[56] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-98)*, 1998.

[57] D. Moore and I. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, April 2003.

[58] D. Gatica-Perez, G. Lathoud, I. McCowan, and J.-M. Odobez, "A mixed-state i-particle filter for multi-camera speaker tracking," in *Proceedings of WOMTEC*, September 2003.

[59] M. Jones and J. Rehg, "Statistical color models with application to skin detection," *Int. J. of Computer Vision*, vol. 46, pp. 81–96, Jan. 2002.

[60] C. Stauffer, "Adaptive Background Mixture Models for Real-Time Tracking," in *Proc. IEEE CVPR*, (Ft. Collins, CO), pp. 246–252, 1999.

[61] R. Collobert, S. Bengio, and J. Mariéthoz, "Torch: a modular machine learning software library," IDIAP-RR 46, IDIAP, Martigny, Switzerland, 2002.

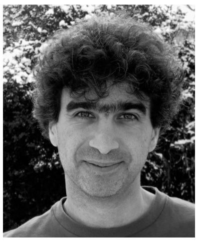
[62] http://www.torch.ch/.

[63] D. Gatica-Perez, I. McCowan, M. Barnard, S. Bengio, and H. Bourlard, "On automatic annotation of meeting databases," in *Proceedings of the International Conference on Image Processing (ICIP 2003)*, 2003.

[64] http://www.itl.nist.gov/div898/handbook/prc/section3/prc33.htm.

Iain McCowan (M'97) received the B.E. and B.InfoTech. from the Queensland University of Technology (QUT), Brisbane, in 1996. In 2001, he completed his PhD with the Research Concentration in Speech, Audio and Video Technology at QUT. He joined the IDIAP Research Institute in April 2001, where he is currently a senior researcher. His research interests include microphone array speech enhancement, audio-visual speaker localisation and tracking, multimodal event recognition, and the application of speech processing to multimedia information retrieval.

Daniel Gatica-Perez (S'01, M'02) received the B.S. degree in Electronic Engineering from the University of Puebla, Mexico in 1993, the M.S. degree in Electrical Engineering from the National University of Mexico in 1996, and the Ph.D. degree in Electrical Engineering from the University of Washington, Seattle, in 2001. He joined the IDIAP Research Institute in January 2002, where he is now a senior researcher. His interests include computer vision, multimodal signal processing, and multimedia information retrieval.



Samy Bengio (M'00) is a senior researcher in statistical machine learning at IDIAP Research Institute since 1999, where he supervises PhD students and postdoctoral fellows working on many areas of machine learning such as support vector machines, time series prediction, mixture models, large-scale problems, speech recognition, multi-modal (face and voice) person authentication, (asynchronous) sequence processing, brain computer interfaces, text mining, and many more.

He has obtained his PhD in computer science from Universite de Montreal (1993), and spent three post-doctoral years at CNET, the research center of France Telecom, and INRS-Telecommunications (Montreal). He then worked as a researcher for CIRANO, an economic and financial academic research center, applying learning algorithms to finance. Before joining IDIAP, he was also research director at Microcell Labs, a private research center in mobile telecommunications. His current interests include all theoretical and applied aspects of learning algorithms.

Guillaume Lathoud (S'02) earned his M.Sc. in Computer Science and Telecommunications in 1999 at Institut National des Telecommunications (INT), France. He then spent more than 2 years as a member of the Digital Television team at National Institute of Standards and Technology (NIST), USA. He joined IDIAP Research Institute, Switzerland in 2002 as a PhD student. His interests include microphone array processing, audio source localization, speaker tracking and multimodal processing.

Mark Barnard (S'02) completed a Bachelor of Computing and Mathematical Sciences with honours at the University of Western Australia in 2000. Since 2001 he has worked as a Research Assistant and PhD student at the IDIAP Research Institute in Martigny Switzerland. His main area of research is in the detection and recognition of events in multimodal data sequences, currently this is focused on event detection for annotation of sporting videos.

Dong Zhang obtained his B.E. (Automatic Control and Electrical Engineering) from Beijing Institute of Technology in 1998, and a M.S. (Computer Vision and Pattern Recognition) from the Institute of Automation, Chinese Academy of Sciences in 2001. Following this, he joined Microsoft Research Asia, where he did research and development in the area of multimedia retrieval. Since 2003 he has been PhD student at the IDIAP Research Institute in Switzerland. His research interests include machine learning, computer vision and multimedia systems.